CLAIMS

What is claimed is:

1. A method for searching for people and organization information on Web pages in a global computer network comprising the steps of:

accessing a Web site of potential interest, the Web site having a plurality of Web pages;

determining a subset of the plurality of Web pages to process; and

for each Web page in the subset, (i) determining types of contents found on the Web page, and (ii) based on the determined content types, enabling extraction of people and organization information from the Web page.

2. A method as claimed in Claim 1 wherein the step of determining content types of Web pages includes obtaining the content owner name of the Web site as a whole by using a Bayesian Network and appropriate tests.

3. A method as claimed in Claim 1 wherein the step of determining content types of Web pages includes collecting external links that point to other domains and extracting new domain URLs which are added to a domain database.

4. A method as claimed in Claim 1 wherein the step of determining the subset of Web pages to process includes processing a listing of internal links and selecting from remaining internal links as a function of keywords.

5. A method as claimed in Claim 4 wherein the step of determining a subset of Web pages to process includes:

extracting from a script a quoted phrase ending in ".ASP", ".HTM" or ".HTML"; and

treating the extracted phrase as an internal link.

6. A method as claimed in Claim 1 wherein the step of determining the subset of Web pages to process includes determining if a subject Web page contains a listing of press releases, and if so, following each internal link in the listing of press releases.

7. A method as claimed in Claim 1 wherein the step of determining the subset of Web pages to process includes determining if a subject Web page contains a listing of news articles, and if so, following each internal link in the listing of news articles.

8. A method as claimed in Claim 1 wherein the step of accessing includes determining whether the Web site has previously been accessed for searching for people and organization information.

9. A method as claimed in Claim 8 wherein the step of determining whether the Web site has previously been accessed includes:

obtaining a unique identifier for the Web site; and

comparing the unique identifier to identifiers of past accessed Web sites to determine duplication of accessing a same Web site.

10. A method as claimed in Claim 9 wherein the step of obtaining a unique identifier includes forming a signature as a function of home page of the Web site.

11. A method as claimed in Claim 1 further comprising imposing a time limit for processing a Web site.

12. A method as claimed in Claim 1 further comprising imposing a time limit for processing a Web page.

13. A method as claimed in Claim 1 further comprising the step of maintaining a domain database storing for each Web site indications of:

Web site domain URL;
name of content owner;
site type of the Web site;
frequency at which to access the Web site for processing;
date of last accessing and processing;
outcome of last processing;
number of Web pages processed; and
number of data items found in last processing.

14 Apparatus for searching for people and organization information on Web pages in a global computer network comprising:

a domain database storing respective domain names of Web sites of potential interest; and

computer processing means coupled to the domain database, the computer processing means:

(a) obtaining from the domain database, domain name of a Web site of potential interest and accessing the Web site, the Web site having a plurality of Web pages;

(b) determining a subset of the plurality of Web pages to process; and

(c) for each Web page in the subset, the computer processing means (i) determining types of contents found on the Web page, and (ii) based on the determined content types, enabling extraction of people and organization information from the Web page.

15. Apparatus as claimed in Claim 14 wherein the computer processing means determining content types of Web pages includes collecting external links and other domain names, and

the step of obtaining domain names includes receiving the collected external links and other domain names from the step of determining content types.

16. Apparatus as claimed in Claim 14 wherein the computer processing means determining the subset of Web pages to process includes processing a listing of internal links and selecting from remaining internal links as a function of keywords.

17. Apparatus as claimed in Claim 16 wherein the computer processing means determining a subset of Web pages to process includes:

extracting from a script a quoted phrase ending in ".ASP", ".HTM" or ".HTML"; and

treating the extracted phrase as an internal link.

18. Apparatus as claimed in Claim 14 wherein the computer processing means determining the subset of Web pages to process includes determining if a subject Web page contains a listing of press releases, and if so, following each internal link in the listing of press releases.

19. Apparatus as claimed in Claim 14 wherein the computer processing means determining the subset of Web pages to process includes determining if a subject Web page contains a listing of news articles, and if so, following each internal link in the listing of news articles.

20. Apparatus as claimed in Claim 14 wherein the computer processing means accessing the Web site includes determining whether the Web site has previously been accessed for searching for people and organization information.

21. Apparatus as claimed in Claim 20 wherein the computer processing means determining whether the Web site has previously been accessed includes:

obtaining a unique identifier for the Web site; and

comparing the unique identifier to identifiers of past accessed Web sites to determine duplication of accessing a same Web site.

22. Apparatus as claimed in Claim 21 wherein the computer processing means obtaining a unique identifier includes forming a signature as a function of home page of the Web site.

23. Apparatus as claimed in Claim 14 further comprising a time limit by which the computer processing means processes a Web site.

24. Apparatus as claimed in Claim 14 further comprising a time limit by which the computer processing means processes a Web page.

25. Apparatus as claimed in Claim 14 wherein the domain database further stores for each Web site indications of:

name of content owner,

site type of the Web site,

frequency at which to access the Web site for processing,

date of last accessing and processing,

outcome of last processing,

number of Web pages processed, and

number of data items found in last processing.